# SPEAKER LOCALIZATION FOR HUMAN-MACHINE INTERACTION

MUHAMMAD AZLAN BIN AHMAD

A report submitted in partial fulfillment of the requirements for the award of the Bachelor of Engineering (Electrical and Electronic)

Faculty of Electrical and Electronic Engineering
University Malaysia Pahang

JULY 2012

# ABSTRACT

Speaker localization is the study direction of sound detection. A listener can identify the location or origin of sound. Human auditory system mechanisms of speaker localization have been widely studied. To adapt the ability of the human auditory system, the researcher uses several methods such as time and level differences between the ears, spectral information, timing analysis, correlation analysis, and pattern matching. The application of the system is implemented to the robot as a model human auditory system. The robot now can interact naturally when human speak to them as it can determine the localization of the speaker. In this paper is to build speaker localization where to determine the sound source in front side of 180 degrees. The azimuth, distance, and also height will be fixed. The sensor is built from the three microphones to detect sound. The speaker source that came will be comparing its intensity at the right and the left side. So, if the source is from the right side, then the servo motor will rotate in the right until the different intensity of sound is same to the both sides. Therefore, the servo motor will stop here. Whereas, this will show the direction of the sound source. The result in the experiment shows that it's quite satisfying with error accuracy as this technique need a more sensitive instrument to compare the sound source.

# ABSTRAK

Penentuan arah bunyi merupakan kajian pengesanan bunyi. Sebagai pendengar boleh mengenal pasti lokasi atau asal bunyi yang dikesan. Mekanisme sistem pendengaran manusia, iaitu penentuan arah bunyi pembesar suara telah dikaji secara meluas. Untuk menyesuaikan keupayaan sistem pendengaran manusia, penyelidik menggunakan beberapa kaedah seperti masa dan tahap perbezaan antara telinga, maklumat spektrum, analisis masa, analisis korelasi, dan pola yang sama. Penggunaan sistem dilaksanakan untuk robot sebagai satu model sistem pendengaran manusia. Robot kini boleh berinteraksi secara semula jadi apabila manusia bercakap dengan mereka kerana ia boleh menentukan arah bunyi. Dalam kertas kerja ini, adalah untuk membina pengesan arah bunyi di mana untuk menentukan sumber suara di sebelah hadapan sebanyak 180 darjah. Pengesan dibina dari tiga mikrofon untuk mengesan bunyi. Sumber pembesar suara yang datang akan dibandingkan keamatan di sebelah kanan dan sebelah kiri. Jadi, jika sumbernya adalah dari sebelah kanan, maka motor servo akan berputar di sebelah kanan sehingga keamatan yang berbeza bunyi yang sama untuk bahagian kedua-dua. Oleh itu, motor servo akan berhenti di sini. Di mana ini menunjukkan arah sumber bunyi. Hasil dalam eksperimen ini menunjukkan bahawa keputusan yang memuaskan dengan sedikit kesilapan ini memerlukan instrumen lebih sensitif untuk membezakan sumber bunyi.

# TABLE OF CONTENTS

# LIST OF FIGURES

**TABLE** **TITLE** **PAGE**

# CHAPTER 1

# INTRODUCTION

## 1.1 Background

Speaker localization is important for human-machine interaction, where the direction of the sound / speaker can be identified. For example, distance-talking speech recognition that detects the direction of voice and microphone signal will be activated to capture clear voice, and also robots require sound localization to seek out or to talk to human beings naturally are the sample of application for speech source localization. When we hear a sound coming from our right, the sound waves reach the right ear first and only then the left ear, which is in the auditory "shadow" cast by our head (head shadow). At the same time, the sound is heard louder in the right ear and its pitch and frequency range is perceived differently on the right side compared to the left side. In technical terms, this is referred to as the interaural (between the ears) time, loudness and frequency difference. We make use of these slight differences to localize sounds and their sources. Therefore, we hear the signal in the right ear earlier and louder than in the left one; this is the cue that the sound source is located to our right. In this paper, a design of a speaker localization system is proposed with two separated channel microphones. First, by comparing the signal, direction of the sound can be determined. In this regard, the geometry of the room and the sensor set-up plays an important role in deriving an accurate speaker position from the Iteraural Loudness Difference estimates.

## 1.2 Problem Statement

Nowadays, human-machine interaction has developed widely as distance speech is one of the kinds. Such as, robot need to capture clear voice of the speaker, therefore, sound localization should be determined.

.

## 1.3 Objective

I. To study about speaker localization method.
II. To achieve sound loudness.
III. To build speaker localization prototype.

## 1.4 Scope

I. Theory and method of speaker localization explain briefly.
II. Three of microphone arrays will be used as a sensor to detect sound in multiple directions on the front side in 2D for 0 to 180 degrees.
III. The azimuth, distance and height of the sound will be fixed.

## 1.5 Thesis Outline

Speaker Localization for Human-Machine Interaction draft thesis consists of three chapters that explain different part of the project. Each chapter elaborates all parts of hardware and programming about this project. The content also consists of information about the project and the component used to as illustrate in the literature review.

Chapter 1 explains the introduction about this research where all the objectives and problems that lead to the implementation of this research are stated. The chapter starts with general information of sound localization and the background of the application.

Chapter 2 explains the literature review that had been studied regarding the speaker localization system project based on recent journals, papers and articles. The information also comes from a few of resources in the internet that can be trusted. But, most of the resources were coming from the e-proxy web of University Malaysia Pahang. Generally, most of the literatures were discussed about the project module from the basic concept to its application to this project and engineering fields.

Chapter 3 will be more focused on methodology of hardware and programming used in the speaker localization system project. Each module has its own connection and condition which needs to put into consideration during the hardware installation. This chapter also explains all the main circuit for each component in more detail.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Introduction

The purpose of this literature review is to identify and synthesize appropriate references to demonstrate and illustrate the presence and absence of knowledge and information regarding speaker localization. This literature review also is an evidence to support the research topic to make easier to understand and to complete the project later on. The journal, conference papers, technical report and other useful resources will be included by summarizing in this literature review section.

## 2.2 Speaker Localization

### 2.2.1 Interaural Intensity Difference

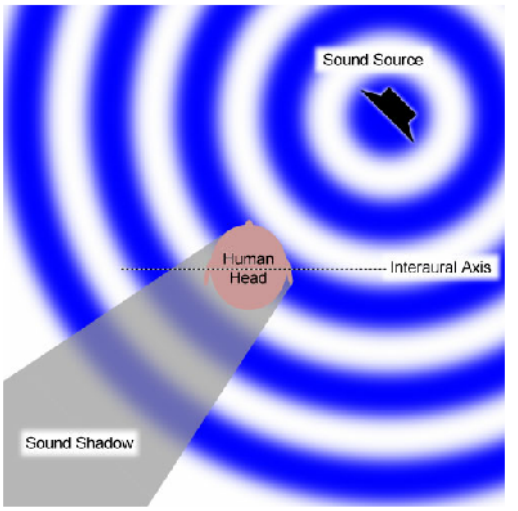


Figure 2.1 Iteraural Loudness Difference

According to the previous studies, one of the cues known to be involved in auditory localization is the relative loudness of the sound at each ear. This is known as the Interaural Loudness Difference (ILD) cue. If a sound comes from directly in front the sound will be exactly the same in both ears. If however, the sound comes from somewhat to the right the sound will be slightly louder in the right ear than in the left. This is due to a sound shadow forming in the far ear due to the sound's direct path being blocked by the head. Low frequencies are less affected by the sound shadow because they tend to bend around large objects. The higher the frequency, the greater the effect of the sound shadow produced by the head. Although the human auditory system makes use of ILDs, ITDs are thought to play a stronger role in how we determine how far left or right a sound source may be. [1]. The Interaural intensity differences, or differences in loudness at the two ears. If the wavelength of sound is less than the diameter of an object in its path, the object creates a "sound shadow" on the side opposite the sound source. This is what our heads do for high frequency sounds, above about 2000 Hz. [2].

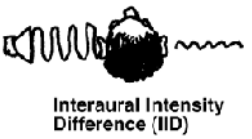


Figure 2.2: Interaural intensity differences are created when the head creates a "sound shadow" on the side opposite the sound source. [2]

### 2.2.2 Shadow Effect in Frequency

The relative loudness is fairly straightforward. If one ear is in the direction of the sound, and the other ear is on the far side of the head from the sound, then the sound shadow" of the head will make the sound softer on the far side than on the near side. However, last lecture we saw that sound tends to bend around obstacles. Whether it does so or not depends on the relative sizes of the object and the wavelength of the sound.

The head will give a good shadow provided that,

$$\lambda_{sound} < D_{head}$$

with $D_{head}$ the diameter of your head, which is around 20 cm = 0.2 m. This gives,

$$\lambda_{sound} < 0.2\text{m}$$

$$f_{sound} = \frac{Vsound}{\lambda sound} > \frac{340m/s}{0.2m} = 1700\text{Hz}$$

Of course, this is not an exact statement. The further above this frequency the sound is the more pronounced the head shadow effect will be. Therefore, it will be a pronounced effect at 5000 Hertz, a modest effect at 2000 Hertz, and nearly nonexistent below 1000 Hertz. [3].Generally, humans are considered to use frequency domain cues to estimate the elevation of a sound source. The frequency response varies with respect to the sound source direction as a result of the interference that occurs between the sound wave that enters the auditory canal directly and the sound wave reflected from the pinnae. In particular, spectral peaks and notches produced respectively by constructive and destructive interference contain information regarding the elevation of the sound source, making it possible to estimate the elevation of a sound source by analyzing them.

Figure 2.3: Developed Pinnae

Spectral cues are dependent on the shape of the pinnae. In this chapter, logarithmic-shaped reflectors were used as pinnae (see Fig. 2.3). The pinnae had a depth of 6 (cm) and were made from 0.5 (mm) thick aluminum sheets. Figure 2.4 shows a photograph of experimental device with the pinnae attached. Figures 2.5 and 2.6 show a front view and a side view of the experimental device with the pinnae attached, respectively.

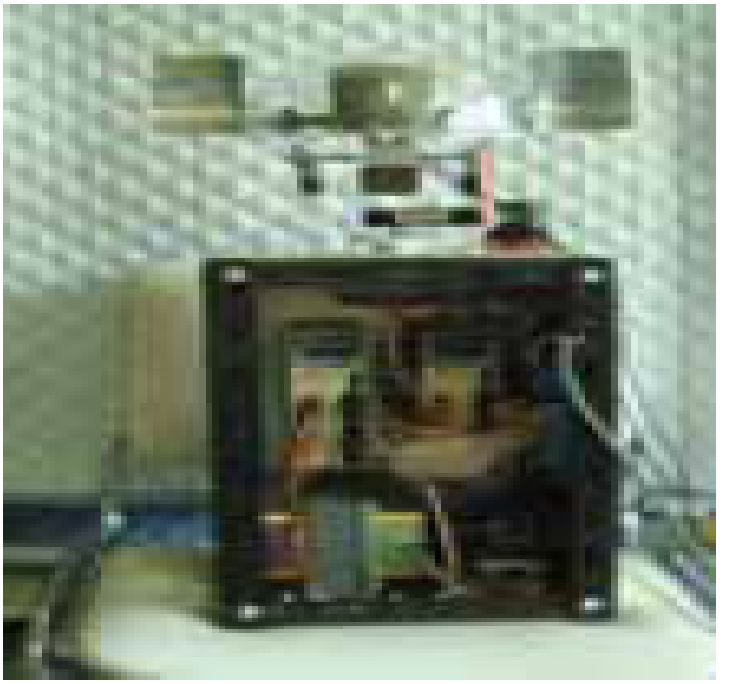

Figure 2.4: Photograph of robot with pinnae attached

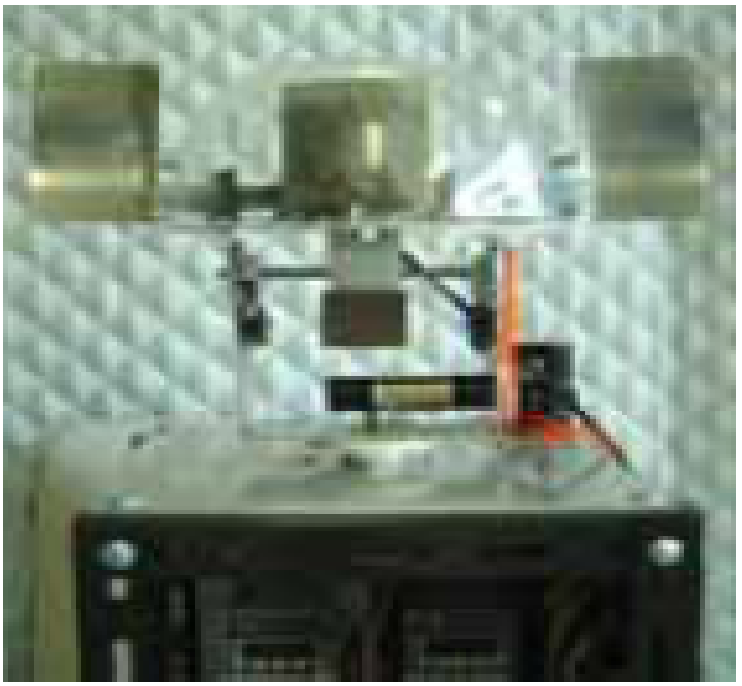

Figure 2.5: Front view of the robot with pinnae

Figure 2.6: Side view of the robot with pinnae

The frequency response of the developed robotic pinna was measured to examine the relationship between spectral cues and sound source elevation. The robot's head was kept still while these measurements were made. A loudspeaker was positioned 0.5 (m) in front of the robot. The frequency characteristics of the pinnae were measured using time-stretched pulses (TSPs). The sound source direction is expressed as

follows. The angle is defined as being 0 (deg) when the sound source is located directly in front of the robot. When the sound source is located below the robot's head, the angle is denoted by a positive value. The results obtained using TSP are shown in Figs. 5(a) to (g). In these results, there are three sharp notches (labeled N1, N2 and N3) within the frequency range from 2 (kHz) to 15 (kHz) and these notches shift to lower frequencies as the robot turned its head upward. Thus, it can be concluded that it is possible to detect the elevation angle of a sound source using pinna cues. [4]

Furthermore, in Interaural Loudness Differences (ILD), with live music, if a violinist is playing a violin in front of you, the loudness at both ears is about equal. If the violinist is standing on your right, the violin sound in your right ear will be louder than in your left. Such loudness differences at the two ears from the same sound source are a cue for the sound's location. ILD cues work well only for signals with energy between 90 Hz and 1,000 Hz. [5]

### 2.2.3 Time Different Of Arrival

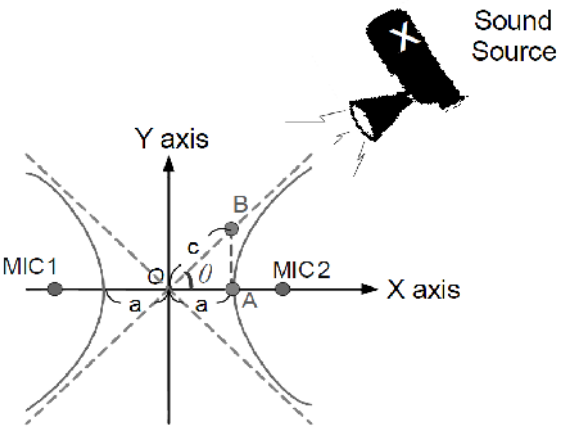


Figure 2.7 Hyperbolic Speaker Location

In Figure 2.7 illustrates the hyperbolic speaker location. To explain the hyperbolic property, by assuming the following assumptions first. Focus-to-focus

distance is 2*c*, namely *MIC1* to *MIC2* distance. The difference between any points on the curve and focus is 2*a*, namely the distance difference between sound and both microphone. Therefore, location of the speech source θ as;

$$\frac{x^2}{a^2} - \frac{y^2}{c^2-a^2} = 1$$

$$2c = |MIC1-MIC2|$$

$$2a = TDOA*v$$

$$\theta = \arccos\frac{a}{c} = \arccos\frac{TDOA*v}{|MIC1-MIC2|}$$

Where *vs* is the velocity of sound.

| Degree / Length | | 0º | 15 º | 30 º | 45 º | 60 º | 75 º |
|---|---|---|---|---|---|---|---|
| 1 meter | Left | 100% | 88% | 100% | 96% | 96% | 88% |
| | Right | | 100% | 100% | 100% | 96% | 100% |
| 2 meter | Left | 92% | 100% | 100% | 100% | 100% | 96% |
| | Right | | 100% | 96% | 96% | 84% | 68% |
| 3 meter | Left | 100% | 100% | 100% | 100% | 92%% | 88% |
| | Right | | 100% | 100% | 100% | 100% | 84% |
| 4 meter | Left | 100% | 100% | 100% | 100% | 88% | 68% |
| | Right | | 100% | 96% | 100% | 88% | 72% |
| 5 meter | Left | 100% | 100% | 100% | 96% | 68% | 60% |
| | Right | | 100% | 96% | 100% | 92% | 56% |

Table 1: All kind of angles accuracy of sound sources localization experimental result. [6]

$$\tau_{ij} = \tau_j - \tau_i$$

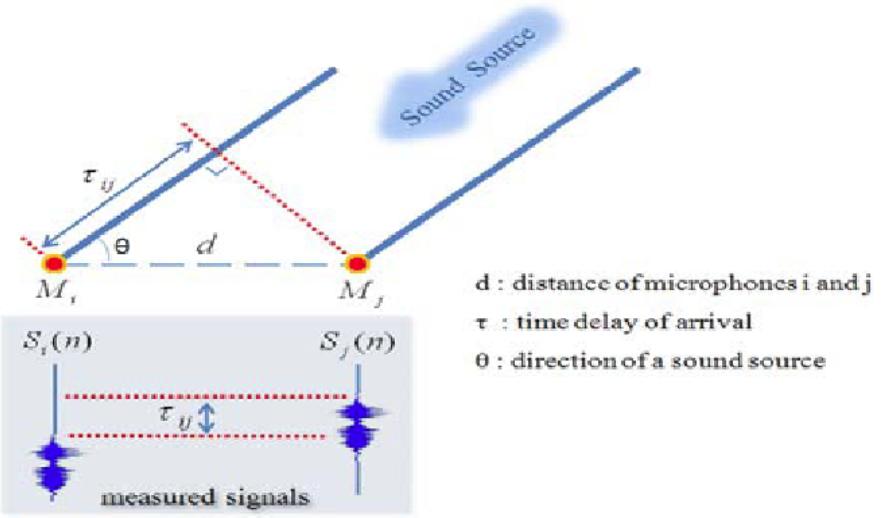


Fig. 2.8: Estimation of a TDOA with a pair of microphones.

The TDOAs obtained from each pair of microphones can be employed by various methods for sound source localization. A method is propose using extracted TDOAs to find the direction of a sound source in the whole azimuth and the height divided into three parts. Since sound source localization systems integrated with a visual system have the horizontal viewing angle of a normal camera, usually more than 10 degrees, then 10 degrees is taken as the unit of azimuth resolution for sound source localization. If the TDOA *τij* is obtained, then the angle $\theta$ characterizing the direction of a sound source in Fig. 2.8 can be estimated. The angle $\theta$ is derived from the following equation.

$$\theta = \cos^{-1}\frac{\tau ij\, C}{d}$$

where *c* is the velocity of sound (340.5 m/s, at 15 °*C*, in air).

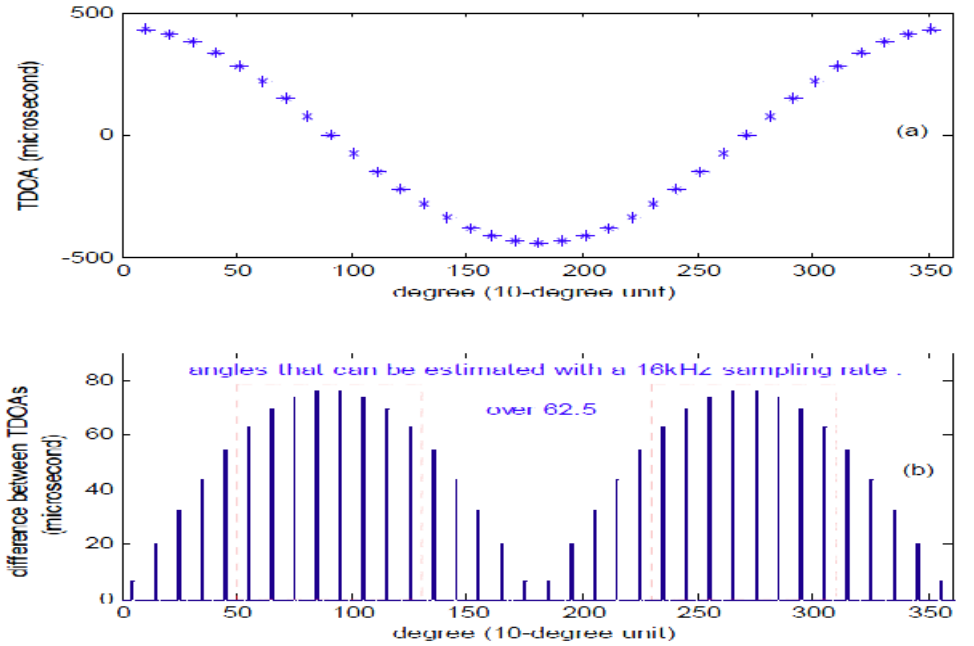


Figure 2.9: (a) Distribution chart of the TDOAs. (b) Differences between TDOAs in 10-degree unit intervals.

Some problems can be identified with finding the direction of a sound source in Fig. 2.9. The results in Fig. 2.9 were simulated with a distance of 15 cm between two microphones and with a 10-degree interval between angles. (a) is the variation of TDOAs driven by angle $\theta$ characterizing the direction of a sound source at 10-degree unit intervals. The TDOAs' variation is an obvious distinction on each angle and it is symmetrical. (b) is a chart of the differences between TDOAs in 10-degree unit intervals. Consider that there is one pair of microphones 15 cm apart and that need to estimate the direction of a sound source using CSP analysis with a sampling rate of 16 kHz and the resolution of a 10-degree unit.

| Azimuth 0º-360º | Height | | |
|---|---|---|---|
| | Over +12cm | Near 0cm | Below -12cm |
| 97.27% | 94.25% | 99.61% | 95.63% |

Table 2: Success Rates of Experiment

The direction of a sound source can only estimate between 50 degrees and 130 degrees without considering the front (0 degree to 180 degree) or back (180 degree to 360 degree) because the TDOA difference that can be estimate is restricted by the sampling rate, and the TDOAs variation also has an angle of symmetry in Fig. 5. The simple solution to problems like these is to extend the distance of the two microphones or to increase the sampling rate, but these also have their limitations, as a matter of course. [7]

In the estimation of speaker direction using cross correlation, by considering the speech signal received by two microphones which are placed at a distance of *d* as shown in Fig. 2.10.

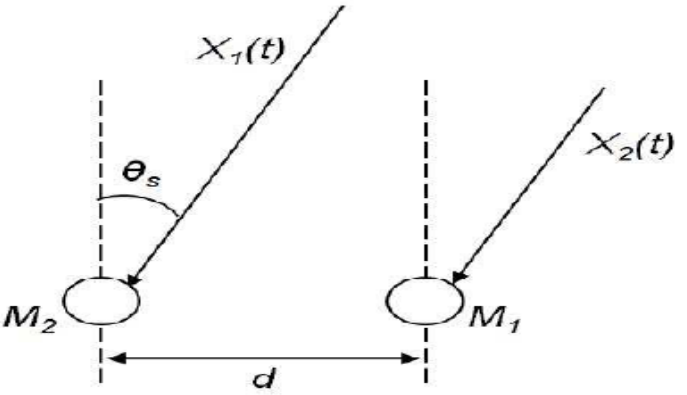


Figure 2.10: Direction of Arrival

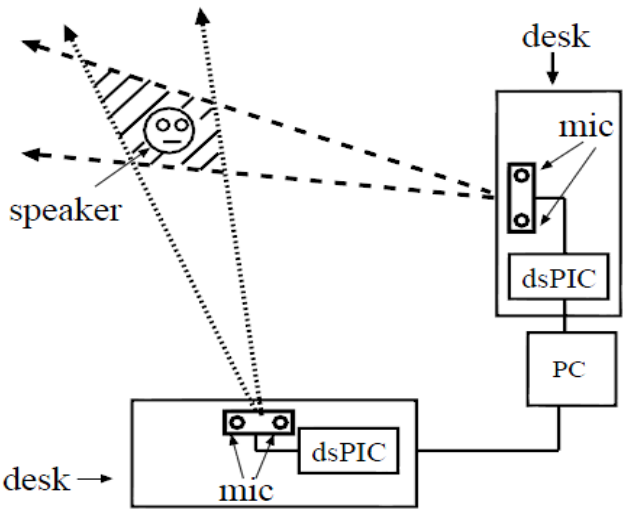


Figure 2.11: Estimation method of speaker location

The signal from the direction θ*s* arrives at the microphone *M*1, then travels a distance *ξ* and arrives at the microphone *M*2. Since an equation $\xi = d \sin \theta s$ is hold, the DOA θ*s* can be calculated by the following equation.

$$\theta s = \sin^{-1} \frac{VTs}{d}$$

where *v* is the velocity of sound and *τs* is a time that the signal requires to travel the distance *ξ*. Therefore, if a correct *τs* (= ^*τ* ) can be obtained, then DOA can be estimated. Many methods for calculating this ^*τ* have been proposed so far. The used approach is TDOA (Time Difference of Arrival), which is one of the most basic and simplest approach, in this paper. In TDOA, ^*τ* can be obtained by maximizing the cross correlation between *X*1(*t*) and *X*2(*t*) as follows [8].

$$\Phi(\tau) = \frac{1}{N} \sum_{t=0}^{N-1} \quad X1(t)X2(t+\tau)$$

Provided that the range of $-T \leq \tau \leq T^{l}$

Furthermore, signal emanating from a remote source and monitored in the presence of noise at two spatially separated sensors can be mathematically modeled as

$$X_1(t) = S_1(t) + n_1(t)$$
$$X_2(t) = \alpha S_1(t + D) + n_2(t),$$

where s1(t), n1(t), and n2(t) are real, jointly stationary random processes. Signal s1(t) is assumed to be uncorrelated with noise n1 (t) and n2(t).

There are many applications in which it is of interest to estimate the delay ***D.*** This paper proposes a maximum likelihood (ML) estimator and compares it with other similar techniques. While the model of the physical phenomena' presumes stationarity, the techniques to be developed herein are usually employed in slowly varying environments where the characteristics of the signal and noise remain

stationary only for finite observation time *T*. Further, the delay *D* and attenuation *a* may also change slowly. The estimator is, therefore, constrained to operate on observations of a finite duration. Another important consideration in estimator design is the available amount of a priori knowledge of the signal and noise statistics. In many problems, this information is negligible. For example, in passive detection, unlike the usual communications problems, the source spectrum is unknown or only known approximately. One common method of determining the time delay D and, hence, the arrival angle relative to the sensor is to compute the cross correlation function.

$$R_{x1x2}(\tau) = E[x_1(t)x_2(t - \tau)]$$

where *E* denotes expectation. The argument $\boldsymbol{\tau}$ that maximizes provides an estimate of delay. Because of the finite observation time, however, Rx1x2(τ) can only be estimated.[9]

The performance of two efficient algorithms for time delay estimation (TDE) has been analyzed. It has been shown that both algorithms are unbiased. Expressions for the time delay estimation mean-square error (MSE) have been presented for both algorithms. It has been shown that the MSE for the algorithm of depends on the unknown delay *D*. Next, both algorithms were combined with the GCC method. It has been shown that they are unbiased, and general expressions for the mean-square error have been presented. It has been shown that a suboptimal estimator for the algorithm of is to use the weight function corresponding to the maximum likelihood estimator (MLE). An optimal estimator (the MSE coincides with the Cramer-Rao lower bound) for the algorithm of is to use the weight function corresponding to the MLE multiplied by $2\pi f$. In this paper, only the local variations in the neighborhood of the unknown delay were studied. The performance of any estimate of $\boldsymbol{D}$ is not fully characterized by the local variations, and the occurrence of false peaks must be taken into account. [10]

Moreover, to improve the performance of TDOA-based 3D localization system in a single-speaker scenario, the proposed modifications were MPHAT (instead of PHAT), a hybrid localization method, and TDOA outlier removal. The GCC-MPHAT method modifies the PHAT weighting function based on an idea borrowed from the generalized spectral subtraction method. The GCC-MPHAT has the advantages of the PHAT method, while it is also robust against noise. In the

hybrid algorithm, the primary estimation of the source location is used to modify erroneous TDOA estimates and find true delays. Consequently, a more accurate estimate of source location is achieved. At the TDOA outlier removal stage, by finding erroneous TDOAs and remove them from the source localization process. These extensive experiments on both simulated and real (practical) data have demonstrated the capability of the proposed modifications in improvement of a speech source localization system. [11]. One of the most important methods of acoustic source localization is the cross-correlation function between two microphones. In general, a higher level (usually the maximum) of correlation means that the argument is relatively close to the actual time difference of arrival (TDOA) between the two microphones for the source. [12]. A circular arrangement of microphones forming an array was studied in makes arrangement is advantageous when the array is used in beam-forming applications to locate the sound source and where the beam is electronically steered in a direction where the speaker is expected to be located. However, this method generally requires a precise calibration of the gain settings of individual microphones.[13]

For the conclusion, the best method in speaker localization is TDOA with correlation, but it is complicated and more complex method. But, the easiest method is IID that will be chosen on this paper. But, IID is the lack of accuracy than TDOA. The error is acceptable and reasonable also really cheap to do this method.

# CHAPTER 3

# METHODOLOGY

## 3.1 Introduction

This section will discuss about the methodology for the feature extraction, evaluation method and the driver condition classification. In previous studies, there is a lot of method to determine the speaker localization. But in this studies, the chosen method will be use is the Interaural Intensity Difference (IID) where the difference in intensity (level) between a sound arriving at one side versus the other side. For example, if the sound arriving at the right side of the sensor, which is this shows at the right side will received the strongest sound signal. But, the left side of the sensor will received a little weak from the right side. So, by comparing the loudness of intensity of the sound, localization of the sound source can be determined. Whereas, the servo motor as the indicator will rotate if there a different between both side of the sensor and it will stop when there is no different in loudness which this shows the direction of the incoming sound.